CQA コンテンツからの類似する悩みの発見

橋口　友哉† 　山本　岳洋†† 　藤田　澄男††† 　大島　裕明†,††

† 兵庫県立大学 応用情報科学研究科　〒 650–0047 兵庫県神戸市中央区港島南町 7-1-28
†† 兵庫県立大学 社会情報科学部　〒 651–2197 神戸市西区学園西町 8-2-1
††† ヤフー株式会社　〒 102–8282 東京都千代田区紀尾井町 1-3
E-mail: †{aa19j508,ohshima}@ai.u-hyogo.ac.jp, ††t.yamamoto@sis.u-hyogo.ac.jp, †††sufujita@yahoo-corp.jp

あらまし　本研究では，悩み相談者が投稿した悩みを含む文をクエリとして受け取り，コミュニティ型質問応答（CQA）コンテンツからこの悩み相談者が共感できると考えられる質問を検索する問題に取り組む．本研究では，悩み相談者が共感できる悩みとは悩みの状況が類似するような悩みであるという仮説を立て，悩みの状況に着目した類似文検索手法を提案する．具体的には,「主人は育児に非協力的で、自分は働いているのだから、家の仕事は私がしろというタイプで、私一人に家事と子育てを任せてくる。」という文と,「旦那は仕事が忙しく、家にあまりいない状況の中、最近は「仕事しないで家にいるんだから子育てくらいまともにしろよ。」と言われました。」という文ペアは状況が類似しているため，同じ悩みを表す類似文として考えられる．類似文検索のためのモデルを構築するため，まず，クラウドソーシングを用い，CQA コンテンツから抽出した文と状況が類似する文を CQA コンテンツから人手で抽出する．次に，得られたデータを用いて BERT のファインチューニングを行い，状況に着目した類似文検索モデルを構築する．得られたモデルの有効性を検証するため，与えられた悩みを含む文から悩み相談者が共感できると考えられる質問を CQA コンテンツから検索する実験を行った．20 件の悩み文をクエリとして実験した結果，TF-IDF，Okapi BM25，事前学習のみの BERT といった手法と比較し，状況の類似性に着目した類似文判定のファインチューニングを行った BERT が最も高い nDCG@1，nDCG@3 を達成した．

キーワード　類似文検索，情報検索，BERT

1 はじめに

Yahoo!知恵袋 [1]や教えて!goo[2]に代表されるコミュニティ型質問応答（CQA）コンテンツでは，質問者は知らない事柄に関する質問をするだけでなく，質問者が抱える悩みの相談がなされることも多い．そのような相談に対して回答者は，その悩みの解決方法を示すだけでなく，回答者が質問者と似たような悩みを抱えている，あるいは抱えていたことを示すことも少なくない．

悩みを持っている人に同じ悩みを持っている人がいると教えることは重要である．同じ悩みや病気を抱える人同士で支え合うピア・サポートは，アルコール依存症や禁煙などのコミュニティで活用されている．中でも，Alcoholics Anonymous は、アルコール依存症者の自助グループとして知られている [3][4]．ピア・サポートについて小野ら [12] は，同じ悩みを持っている人がいることを知ることは重要であると報告している．さらに,「患者にとって本当の意味での共感はピア（同じ立場の人間）だからこそ成しえるものであり，非常に重要なサポートである」と述べている．

本研究では，CQA コンテンツにおいて質問者が投稿した悩みを対象に，その悩み相談者が共感できるような類似する悩みを含む質問を CQA コンテンツにアーカイブされた質問から発見する技術について取り組む．類似する悩みを発見することが可能となれば，CQA コンテンツで質問を探す際に，質問者が類似する悩みを発見する助けになる．また，システムがユーザの悩みを捉えることが出来るため，ユーザの悩みに寄り添う悩み対話システムを構築することも出来ると考えられる．

本研究では産後うつに関する悩みを対象とする． 産後うつは産後間もない母親が育児によるストレスや周囲の環境等によって発症する症状である．また，産後うつは他者に自身の心境を話して，理解してもらうことが治療の上で大事だといわれている．岡野ら [9] によると，周産期になんらかの精神健康上の問題により苦悩する女性が少なくないことも知られている．玉木ら [10] によると，産後間もない母親は産後に精神的不調を感じても，専門家や専門機関に相談しないと報告している．また，健やか親子 21[5]によると，産後のメンタルヘルスについて議論されている．そのため，産後の悩みを抱える相談者にシステムが類似する悩みを提示することは，効果的であると考えられる．

本研究では類似する悩みを発見するため，悩みの状況に着目したモデルの作成を提案する．本研究で着目した悩みの状況は具体的には,「主人は育児に非協力的で、自分は働いているのだ

1 : https://chiebukuro.yahoo.co.jp/
2 : https://oshiete.goo.ne.jp/
3 : https://aajapan.org/
4 : https://aa.org/
5 : http://sukoyaka21.jp/about，2020 年 1 月 1 日　閲覧

から、家の仕事は私がしろというタイプで、私一人に家事と子育てを任せてくる」だと「悩みの状況」は「主人は育児に非協力的」と「私一人に家事と子育てを任せてくる」となる．このように悩みには悩み相談者の状況が記述されている場合がほとんどである．しかし，従来の類似文検索手法では，状況を捉えず，類似性評価を行うため，真に類似する悩みを上手く発見することができないと考えられる．たとえば，TF-IDF や Okapi BM25 による手法は単語の有無に基づき類似性を評価するため，「姑が嫌い」と「義母を憎んでいる」の類似性を捉えることができない．また，深層学習を用いた汎用言語モデルである BERT でも文全体の意味で類似性を評価するため，悩み相談のような自由投稿では個人の記述の仕方によって，文としての類似性を正しく捉えられない可能性がある．本研究では，悩みの類似性は状況が重要であると仮定し，状況に着目したモデルを作成することで検証した．

本研究の提案手法である類似する悩みを自動発見するためのアプローチについて説明する．まず，モデル構築のため，データセットをクラウドソーシングにより収集した．具体的には，Yahoo!知恵袋から抽出した産後うつに関する悩みの文をワーカに与え，その悩みと状況が類似する文を探し出してもらうタスクにより悩みの状況が類似する文を収集する．また，収集後に「悩みの状況」にあたる単語をアノテーションしてもらうことで，悩みの状況に着目したデータセットの収集も行った．次に，クラウドソーシングで集めたデータセットを用いて，汎用言語モデルである BERT をファインチューニングした．BERT のファインチューニングは，2 種類の手法で行った．1 つ目は，クラウドソーシングのアノテーション結果を用いて与えられた文から状況にあたる単語を推定するタスクである．2 つ目は，与えられた 2 つの文が同一の悩みであるかを判定する 2 クラス分類タスクである．

提案手法の有効性を検証するため，与えられた悩みを含む文から悩み相談者が共感できると考えられる質問を CQA コンテンツから検索する実験を行った．20 件の悩み文に対して評価を行った結果，TF-IDF，Okapi BM25，事前学習のみの BERT といった手法と比較し，状況の類似性に着目した類似文判定のファインチューニングを行った BERT が最も高い nDCG@1，nDCG@3 を達成した．

2 関 連 研 究

医療分野ではピアとは「同じ病気にかかっている，あるいは同じ身体障害をもっている人同士」と規定されている [11]．1 章でも述べたように，ピア・サポートを行うことで，患者同士は精神面や QOL によい影響があるといわれている [7] [8] [12]．Social comparison 理論では，自分と似た経験をもつ他者と自分を比較できることは，その経験を常態化し，肯定的なロールモデルを得られる．結果，ヘルスプロモーション行動を推進させ，自尊心を高めるとされている [3]．

本研究は BERT [4] を用いて類似文を検索する．BERT は Google が提案した汎用言語モデルである．このモデルをタスクに応じてファインチューニングすることで，様々なタスクで精度向上が達成されている．本研究でも BERT のファインチューニングを行うことで，類似文検索の精度向上がみられるかを確認する．MacAvaney ら [5] も BERT を用いて類似文検索を行っている．MacAvaney ら [5] によると，BERT 学習時に使用される [CLS] 特殊トークンを既存の類似文検索ランキングモデルに組み込むことで精度が向上したといわれている．そこで，本研究でも，MacAvaney らと同様に [CLS] 特殊トークンを用いて類似文検索を行う．類似文検索に BERT を利用した他の研究としては以下のものがある [2] [6]．Asai ら [2] はユーザの入力クエリから該当する可能性がある記事を TF-IDF で絞り込んだ後，BERT を用いて記事の選択を行っている．Sakata ら [6] はユーザの入力クエリから類似の質問をとる際に，BERT を入力クエリと回答との関連性を出すために使用している．具体的には，Okapi BM25 により計算される入力クエリと質問の類似性と BERT で得られた入力クエリと回答の関連性を組み合わせることで，質問のランキングを行っている．

Yilmaz ら [1] は BERT の入力長を超える文書に対して，文に分割して BERT へ入力する手法を用いている．Yilmaz らは本研究とは異なり，BERT を用いて，文をベクトル化していない．Yilmaz らはクエリと文の類似確率を用いて，スコアリングし，BM25 のスコアと組み合わせてリランキングしている．

3 クラウドソーシングによる状況が類似する悩みの収集

本章では，クラウドソーシングを用いたデータセット収集について説明する．本クラウドソーシングの目的は，悩みを含むある文に対して類似する悩みを含んだ文を収集することによる，類似文検索のためのデータセット構築，および，4 章で述べる BERT のファインチューニングのためのアノテーション，という 2 点である．

以降，まずクラウドソーシングで収集対象とした，悩みを含む文の選定方法について述べる．その後，実際のクラウドソーシングタスクについて述べた後，得られたデータセットの概要を示す．

3.1 悩みを含む文の収集

本研究は，Yahoo!知恵袋から産後うつに関する悩みを抱えていると思われるカテゴリに投稿された質問から，悩みを含む文を抽出した．対象にしたカテゴリは以下の 10 個である．

- 子育て、出産
- 子育ての悩み
- 妊娠、出産
- 幼児教育、幼稚園、保育園
- 恋愛関係、人間関係の悩み
- 家族関係の悩み
- 病院、検査
- 病気、症状
- 女性の病気

- 生理

これらのカテゴリに含まれる質問から，まず産後うつに関する質問を自動で抽出した．具体的には，質問とその質問に対する回答のどちらかに「産後うつ」,「産後鬱」あるいは「マタニティーブルー」というキーワードが 1 つでも含まれる質問を抽出した．その後，抽出された質問を人手で確認し，産後うつに関する悩みを含んでいる 1 文を抽出した．ここで，文の定義は GiNZA[6]により分割される文を 1 文の単位とした．本研究では，産後うつに関する悩みを含む 50 文を用意した．

3.2 タスクの実施手順

クラウドソーシングで実施したタスクの手順について説明する．クラウドソーシングのタスクは以下の手順で行った．

まず，ワーカに本タスクの目的と概要を説明し，タスクの内容が学術研究に利用されることや報酬などに関する説明ページを提示した．上記に同意したワーカのみ以下に示す実際のタスクに進んだ．

（1） ワーカに，取得して欲しい類似文に関する説明を具体例とともに行った．ワーカに実際に提示した 1 例を示す．

> 探してきていただきたい質問は,「悩みの状況」が似ているような質問です.
> **悩み:** 主人は育児に非協力的で、自分は働いているのだから、家の仕事は私がしろというタイプで、私一人に家事と子育てを任せてくるので、育児と家事の両立に手が回りません。
> **探してきてほしい質問:** 旦那は仕事が忙しく、家にあまりいない状況の中、最近は「仕事しないで家にいるんだから子育てくらいまともにしろよ。」と言われました。

（2） 次に，前節で述べた 50 文から 1 文をワーカに提示し，以下の 3 つのサブタスクを行ってもらった．

（a） Yahoo!知恵袋の検索ページ[7]にアクセスし，自由に検索をしてもらい，与えられた文と類似する悩みを持つ質問を探してもらう．その後，その質問から類似する悩みを含む 1 文を選択してもらう．このとき，検索対象とする質問の期間は，2004 年 4 月 1 日から 2018 年 4 月 24 日に投稿された質問に限定してもらった．

（b） （a）で選択した文の中で，状況をよく表していると思われるフレーズを 1 つ以上記述してもらう．

（c） この質問を探す際に使用した検索クエリを入力してもらう．

（3） 手順（2）を 5 つの文に対して行ってもらった．

（4） 最後に，性別と年代について回答してもらった．

上記タスクを 10 タスク，計 50 文に対して実施した．1 つのタスクに対して 20 名のワーカに回答してもらった．また，ワーカは 1 つのタスクあたり 1 回しか実施できないが，異なるタスクを実施することは可能とした．

6 : https://megagonlabs.github.io/ginza

7 : https://chiebukuro.yahoo.co.jp/search/advanced?fr=common-navi

3.3 収集したデータセットの概要

2019 年 12 月 25 日から 27 日にかけて，前節で述べたタスクをクラウドソーシングプラットフォームであるランサーズ[8]で実施した．タスクを完了したワーカには報酬として 300 円を支払った．クラウドソーシングの結果，66 名のワーカから 50 件の文に対してそれぞれ 20 件，合計 1,000 件の類似文を収集した．

次に，得られた 1,000 件のデータから不適と思われる文を人手で除去した．具体的には，1 文ではなく 2 文以上の文で回答がなされたもの（272 件），ワーカが文を意訳して単語を抽出したり文に含まれない単語を状況として回答したもの（33 件）を除去した．その結果，50 件の文に対して合計 695 件の類似文を得ることができた．以降，本研究で行う実験とモデル構築にはこのデータセットを用いる．得られたデータセットの具体例を表 1 に示す．

4 類似する悩みを含む文発見のための類似文検索手法

本研究では，与えられた悩みを含む文と状況が類似する悩みを含む文を検索するため，近年の自然言語処理タスクで広く利用されている BERT を利用した類似文検索を提案する．さらに，類似する悩みの発見に，より特化したモデルを構築するため，BERT のファインチューニング手法についても提案する．

4.1 BERT

BERT（Bidirectional Encoder Representations from Transformers）[4] は大規模なコーパスで事前学習することで汎用言語モデルを構築する．BERT では Masked LM と Next Sentence Prediction (隣接文予測) の 2 つのタスクを大規模コーパスにより学習することで，言語モデルを構築する．Masked LM はランダムに MASK 化された単語を当てるタスクである．隣接文予測はある 2 つの文が同一ドキュメントかを当てるタスクである．その後，解きたいタスクへモデルをファインチューニングすることで，モデルのパラメータをタスクに最適化させる．BERT はさまざまな自然言語処理タスクに適用され，高い精度を達成している．

本研究では，BERT の日本語の事前学習モデルとして京都大学大学院黒橋・河原研究室が提供するモデル[9]を用いた．本研究での BERT の使用目的は文のベクトル化とファインチューニングである．BERT へ必ず入力される [CLS] 特殊トークンは，事前学習の際に文の関連性判定に使われているため，文を表していると考えられている．本研究では，BERT を用いた類似文検索は BERT に文を入力し，[CLS] 特殊トークンに対応するベクトルを取得することで行う．本研究で指定するレイヤーは最終層の 1 つ前とした．得られたベクトルを文ベクトルとみなし，文間のコサイン類似度をとることで文の類似度を求める．

8 : https://lancers.jp

9 : http://nlp.ist.i.kyoto-u.ac.jp/index.php?BERT 日本語 Pretrained モデル

表 1　クラウドソーシングのタスク実行結果を抜粋.

提示した悩み文	ワーカが出した類似文	状況を表す単語
夜中の授乳や家で赤ちゃんと二人きりの時にとても不安になります。	恥ずかしい話、いざ赤ちゃんと二人きりになった途端寂しさと不安感が込み上げてきました。	赤ちゃんと二人きり、寂しさ、不安感
旦那は育児に非協力的で、私が泣いたり、怒ったり、弱音を吐くと母親失格と激怒します。	旦那は育児に非協力で、どうしても私が面倒みれない時のみは面倒見れますが、「基本的にはお前がやれよ！」ってタイプです。	旦那は育児に非協力、「基本的にはお前がやれよ！」ってタイプです

4.2　BERT のファインチューニング

前節で述べた BERT を用いて文間の類似度を計算することによって，与えられた文と類似する文を発見することが可能となる．しかし，すでに述べたように悩みの類似性では状況の類似性が重要であると考えられる．そのため，与えられた文全体をみて類似度を計算するモデルでは悩みの類似性を正しく捉えられないのではないかと我々は考えた．

そこで，本研究では前節で述べたクラウドソーシングにより収集したデータセットを用い BERT をファインチューニングすることで，より悩みの類似性に特化したモデルを構築することを試みる．本研究では 2 種類のファインチューニング方法を実装した．

それぞれのファインチューニング方法にある [CLS] と [SEP] と [PAD] は BERT での学習時に用いる特殊トークンである．[CLS] は文頭，[SEP] は文末，[PAD] は入力長が 128 単語以下の場合に 128 単語になるようにそれぞれつけられる．

4.2.1　悩み文からの状況推定によるファインチューニング

1 つ目の手法は，悩み文からの状況推定によるファインチューニングである．具体的には,「私は結構感情で怒ってしまいいつも子どもが寝た後反省というか自己嫌悪感で一杯です。」という入力に対し,「感情で怒ってしまい」と「自己嫌悪感で一杯」という「悩みの状況」を表す単語を推定する．このようなタスクを解くことで，悩みの状況に着目したモデルを構築することができると考えられる．また，単語に対してラベルをつけるということから，このタスクは系列ラベリングタスクとして解くことができる．本研究では，状況を推定することで，文全体の意味だけでなく，悩みとして重要な単語に着目した類似文検索を行うことができると考えた．

本研究で悩み文の状況推定に使用した訓練データの例を表 2 に示す．状況推定によるファインチューニングは，[状況]，[非状況]，[CLS]，[SEP]，[PAD] の多クラス分類である．単語につけられている [状況] ラベル，[非状況] ラベルはそれぞれワーカが状況を表していると回答した単語，それ以外の単語である．

悩み文の状況推定に用いたモデルを図 1 に示す．このモデルは各単語に BERT を適用し，ベクトル化した後，出力層の活性化関数にソフトマックス関数を用いることで，単語がどのラベルに属するのかを判定するモデルである．また，誤差関数には多クラス交差エントロピーを用いた．

4.2.2　状況類似判定によるファインチューニング

2 つ目の手法は，悩み文の状況類似判定によるファインチューニングである．隣接文予測タスクは，文が 2 つ与えられたときにそれらが類似する文かを予測する 2 クラス分類タスクであり，BERT の事前学習にも用いられている．このようなタスクを解

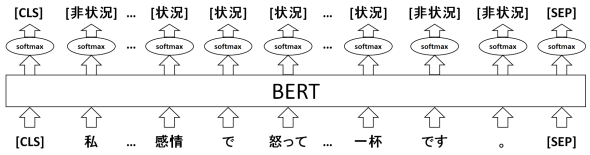

図 1　悩み文の状況推定による BERT ファインチューニングモデル.

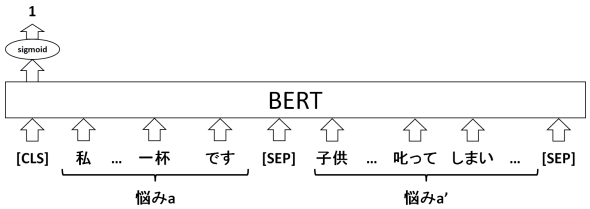

図 2　状況類似判定による BERT ファインチューニングモデル.

くことで，悩みの状況類似判定モデルを構築することができると考えられる．また，状況が類似する悩みを判定することができるので，類似文検索の精度が向上すると考えられる．本研究では，クラウドソーシングのタスクで得られたある悩みと状況が類似すると判断された文を用いて隣接文予測タスクを実行することで悩み文の状況類似判定を行う．

本研究で状況類似判定に使用した訓練データの例を表 3 に示す．状況悩み文判定によるファインチューニングは 2 つの文が類似するかを判定する 2 クラス分類タスクである．類似する文であれば 1，そうでなければ 0 となる．このタスクは [CLS] 特殊トークンで 2 文が類似するかを判定する．また，[SEP] 特殊トークンで文の区切りをとる．学習には類似する悩み文ペアと類似しない悩み文ペアを用意する必要がある．これは，前述したクラウドソーシングで構築したデータセットを用いて用意する．学習に用いる訓練データの詳細は次章の実験設定にて述べる．

状況類似判定に用いたモデルを図 2 に示す．このモデルは入力する 2 文に BERT を適用した後，[CLS] 特殊トークンの出力層の活性化関数にシグモイド関数を用いることで，2 文が類似するかを判定するモデルである．また，誤差関数にはバイナリ交差エントロピーを用いた．

5　評　　価

提案手法の有効性を検証するため実験を行った．まず，実験に用いた手法について説明し，クラウドソーシングで収集したデータセットのみを用いた精度評価について説明する．その後，本研究の目的である，与えられた悩みに対する，CQA コンテンツからの類似質問検索の精度評価について説明する．

表 2　状況推定に用いた訓練データ例.

単語	[CLS]	私	は	結構	感情	で	怒って	...	一杯	です	。	[SEP]	[PAD]
ラベル	[CLS]	[非状況]	[非状況]	[状況]	[状況]	[状況]	[状況]	...	[状況]	[非状況]	[非状況]	[SEP]	[PAD]

表 3　類似悩み文判定に用いた訓練データ例.

入力	ラベル
[CLS] 私は結構感情で怒って... 自己嫌悪感で一杯です。[SEP] 子供が言うことを聞かなくて、カッとなって感情的に叱ってしまい、後で反省…ということが最近続いています。 [SEP]	1
[CLS] 私は結構感情で怒って... 自己嫌悪感で一杯です。[SEP] 旦那は育児を進んでは絶対にやりません。 [SEP]	0

5.1　実験に用いた手法

本研究では，類似する悩み文検索のための手法として，BERT のファインチューニングを提案した．提案手法の有効性を検証するため，以下の手法を実装し，実験に用いた．

5.1.1　TF-IDF

文書検索の伝統的なモデルである，TF-IDF によるベクトル空間モデルを比較手法として用いた．TF-IDF は文に含まれる単語の重要度を算出する手法であり，情報検索の類似文検索でよく使われる．そのため，本研究の比較手法の一つとして用いる．TF-IDF は，同じ単語がある文が類似する手法である．TF-IDF の式は以下になる．

$$\mathrm{tf}(w,s) = \frac{n_{w,s}}{\sum_{w_i \in s} n_{w_i,s}} \tag{1}$$

$$\mathrm{idf}(w) = \log \frac{1+N}{1+\mathrm{df}(w)} + 1 \tag{2}$$

$$\text{tf-idf}(w,d) = \mathrm{tf}(w,s) \cdot \mathrm{idf}(w) \tag{3}$$

ここで，$n_{w,s}$ は文 s 内の単語 w の出現頻度，$\sum_{w_i \in s} n_{w_i,s}$ は文 s 内のすべての単語の出現頻度の和である．また，N は検索対象となる全文数，$\mathrm{df}(w)$ はそのなかで単語 w を含む文の数である．本研究では，質問内の 1 文を文書とみなし，ベクトル空間モデルに基づいて類似文のランキングを行った．すなわち，文を上記 TF-IDF で重み付けした特徴ベクトルを文のベクトルとして用い，入力の悩み文とコサイン類似度の高い文を類似文とみなした．なお，TF-IDF の実装は scikit-learn[10] の TfidfTransformer を用いた．

5.1.2　Okapi BM25

テキストベースの情報検索のベースラインとして良く利用される BM25 についても，本研究の比較手法として用いた．Okapi BM25 の式は以下のように表される．

$$\mathrm{score}(s,q) = \sum_{i=1}^{n} \mathrm{idf}(w_i) \cdot \frac{f(w_i,s)\cdot(k_1+1)}{f(w_i,s) + k_1 \cdot (1 - b + b \cdot \frac{|s|}{\mathrm{avgsl}})} \tag{4}$$

$$\mathrm{idf}(w_i) = \log \frac{N - n(w_i) + 0.5}{n(w_i) + 0.5} \tag{5}$$

ここで，$q = \{w_1 \cdots w_n\}$ は入力として与えられる悩み文であり，$w_1, \cdots, w_n$ は q に含まれる単語集合である．$f(w_i, s)$ は文 s における単語 w_i の出現頻度，$|s|$ は文 s の単語数，$avgsl$ は検索対象の文集合における平均単語数である．また，N は検索対象となる全文数，$n(w_i)$ はその中で単語 w_i を含む文数である．Okapi BM25 のパラメータは $k_1 = 1.2$，$b = 0.75$ とした．また, idf 値の最小値として $\epsilon = 0.25$ 以上の idf 値を持つ単語のみを計算の対象とした．入力悩み文に対して Okapi BM25 のスコアが高い文ほど類似文であるとみなし，ランキングを行う．

5.1.3　BERT（事前学習のみ）

4.1 節で述べた事前学習済みの BERT を用いた手法である．具体的には，BERT に対して入力文を入力し，出力層の [CLS] トークンのベクトルを文の特徴ベクトルとして抽出し，文同士の類似度をコサイン類似度によりランキングする．

5.1.4　BERT（ファインチューニング）

BERT のファインチューニングの実装には TensorFlow（バージョン 1.15），GPU には GeForce RTX 2080Ti(11GB) を用いた．悩み文からの状況推定によるファインチューニングには GitHub に公開されているコード [11] を用いた．

状況類似判定によるファインチューニングには GitHub に公開されているコード [12] を用いた．状況類似判定によるファインチューニングのための訓練データは以下の様に用意した．まず，訓練データに含まれる，悩み文とその悩み文に対して類似悩み文と判定された文ペアを正例とみなした．そして，他の悩み文に対して類似文と判定された文ペアを負例とみなし，正例と同数の負例を訓練データから抽出し学習に用いた．

BERT のファインチューニングに用いたハイパーパラメータは状況推定と状況類似判定どちらも同じものを用いた．本研究で用いたハイパーパラメータを表 4 に示す．また，学習アルゴリズムには Adam を用いた．本研究ではハイパーパラメータのチューニングについては行わないこととし，ハイパーパラメータは既存実装や既存研究を参考に決定した．

5.2　評価指標

本研究では，評価指標として MAP と nDCG を用いる．各評価には NTCIREVAL[13] の AP と nDCG を用いた．検索対象の適合度の判定は次節以降で述べる．

10 : https://github.com/scikit-learn/scikit-learn

11 : https://github.com/kyzhouhzau/BERT-NER

12 : https://github.com/google-research/bert

13 : https://github.com/mpkato/pyNTCIREVAL

表 4　ファインチューニングに用いたハイパーパラメータ.

ハイパーパラメータ	数値
最大トークン数	128
バッチサイズ	32
学習率	2.0×10^{-5}
エポック数	4
warmup proportion	0.1

表 5　クラウドソーシングで構築したデータセットに基づく評価結果.

	MAP	nDCG@1	nDCG@3	nDCG@5
TF-IDF	0.276	0.900	0.800	0.676
Okapi BM25	0.247	0.700	0.617	0.559
BERT(事前学習のみ)	0.100	0.500	0.364	0.306
BERT(状況推定)	0.055	0.300	0.200	0.158
BERT(状況類似判定)	0.227	0.700	0.594	0.528

5.3　クラウドソーシングで構築したデータセットを用いた評価

評価にあたり，まず 3 章で述べた，クラウドソーシングで構築したデータセットのみを用いて，提案手法の評価を行った．この実験の目的は，BERT のファインチューニングを行うことで，事前学習のみの BERT と比べて類似文の判定精度が向上するかどうかを確認することが主な目的である．

構築したデータセットは，クラウドソーシングで用いた 50 件の産後うつに関する悩み文それぞれに対して，ワーカが類似文だと判断した文が複数存在する．ワーカが類似文だと判断した文は，入力となった悩み文に対する類似文であると判断することができる．また，入力となった悩み文以外の悩み文に対してワーカが類似文と判断した文については，入力となった悩み文に対しては類似していない，つまり不適合であると考えることができる．このように仮定を置くことで，クラウドソーシングで収集したデータからある悩み文に対して正解となる類似悩み文，不正解となる類似悩み文を自動的に決めることができる．本節ではこのデータを用いて手法のランキング精度を評価する．

評価にあたり，まずデータセットを訓練データとテストデータに分割した．構築した 50 件の悩み文に関するデータセットから，40 件の悩み文とクラウドソーシングで収集したそれらに対する類似文を訓練データに，残りの 10 件の悩み文とそれらに対する類似文をテストデータとした．この訓練データを用いて，5.1.4 節で述べた 2 種類のファインチューニングをそれぞれ行いモデルを構築した．そして，テストデータとなる 10 件の悩み文に対して，それら 10 件に対する全類似文を検索対象とし，各種手法での類似悩み文のランキング精度を評価した．

テストデータに対するランキング評価結果を表 5 に示す．なお，適合度については，先述した正解となる類似文を適合，不正解となる文を不適合として MAP および nDCG を計算している．表 5 の事前学習のみの BERT，状況推定のファインチューニングを行った BERT，状況類似判定のファインチューニングを行った BERT を比較すると，状況推定を行うファインチューニングは精度向上に寄与していないものの，状況類似判定のファインチューニングを行うことで検索精度が向上していることが分かる．

一方，表 5 より，ベースライン手法である TF-IDF に基づく手法が MAP，nDCG 両者とも最も高い値となっており，続いて Okapi BM25 および状況類似判定のファインチューニングを行った BERT が高い精度となっていることが分かる．TF-IDF と Okapi BM25 の精度が高い理由の 1 つとして，クラウドソーシングで収集する際にキーワード検索を行って類似する悩みを含む質問を探してもらった点があげられる．キーワード検索を用いて類似質問を検索し，そこから類似する悩み文を抽出すると，入力となる悩み文と，類似する悩み文の単語が重複する可能性が非常に高くなると考えられる．実際，クラウドソーシングでワーカが質問検索のために用いた検索クエリで良く用いられた単語を分析したところ，上位 3 件の単語はほぼすべて，提示した入力悩み文に含まれている単語であった．したがって，TF-IDF と Okapi BM25 といった単語ベースの手法が高い精度となったのではないかと考えられる．

5.4　CQA コンテンツからの類似質問検索による評価

本研究の目的は，ある悩み相談者の悩み文を受け取り，その悩み相談者が共感できると考えられる類似質問を検索することであった．そこで，この問題設定に即した実験設定で実験を行い，手法の性能を評価する．具体的には，入力として与えられた悩み文から，その悩み相談者が共感できる考えられる類似質問を Yahoo!知恵袋から検索する，という設定で実験を行う．

5.4.1　クエリとコーパスの用意

まず，この実験にあたりクエリとして用いる悩み文を，3 節で述べた 50 件とは別に新たに 20 件用意した．そして，検索対象とする質問集合を以下の手順で用意した．まず，3.1 節で述べた，産後うつに関する質問から Yahoo!知恵袋から無作為に一定割合抽出し，そこから 3 節で述べたクラウドソーシングでワーカが類似文を抽出した質問を除去した．この結果，3,113 件の質問を抽出し，本節での検索対象とした．

5.4.2　質問ランキング手法

次に，各手法で入力された悩み文から質問をランキングする方法について述べる．本研究では，質問を文に分割し，5.1 節で述べた手法で入力された悩み文と質問中の各文との類似度を計算し，その類似度の最大値に基づいて質問をランキングする．いま，質問 Q が n 個の文 $Q = \{s_1, \cdots, s_n\}$ からなるとき，入力となる悩み文 q に対する Q のスコアを下記の式に従って求める．

$$\mathrm{score}(q, Q) = \max_{s_i \in Q} f(s_i, q) \tag{6}$$

ここで，$f(s_i, q)$ は 5.1 節の述べた，各手法における悩み文と文の類似度計算方法である．

5.4.3　適合度判定

各手法の精度を評価するため，それぞれの悩み文に対してそれぞれの手法で上位 3 件までにランキングされた質問をプーリングし，適合度の判定を行った．適合度の判定についてはクラウドソーシングを使用した．具体的には，クエリとして与えられた悩み文 1 つと，その悩み文の各手法から得られた質問集合をワーカに対してみせ，その質問集合から入力として与えた悩

表 6　CQA コンテンツからの共感できる類似質問検索に基づく評価結果（2 値適合性）

	nDCG@1	nDCG@3
TF-IDF	0.300	0.318
Okapi BM25	0.150	0.134
BERT(事前学習のみ)	0.300	0.382
BERT(状況推定)	0.300	0.265
BERT(状況類似判定)	0.400	0.387

表 7　CQA コンテンツからの共感できる類似質問検索に基づく評価結果（多値適合性）.

	nDCG@1	nDCG@3
TF-IDF	0.250	0.290
Okapi BM25	0.125	0.124
BERT(事前学習のみ)	0.200	0.358
BERT(状況推定)	0.200	0.242
BERT(状況類似判定)	0.250	0.326

み相談者が最も共感できると考えられる質問を 1 つだけ選択する，というタスクを行ってもらった．また，ワーカがタスクを真面目に取り組んでいるかどうかを判断するため，質問を選択する際にその判断の根拠についても記述してもらった．このタスクを 1 つの悩み文あたり 30 人のワーカに対して行ってもらい，それを基に質問の適合度を用意した．

クラウドソーシングで得られるデータは最も共感できると判断された質問とその人数であり，人数をそのまま適合度とする方法は適合度の範囲が 0 から 30 と非常に大きくなるため不適切であると考えた．そこで，本研究では以下の 2 種類の方法で，クラウドソーシングにより得られた 30 人の判定を質問の適合度へと変換した．

- 2 値適合性: 少なくとも 2 人以上に最も共感できると判断された質問を適合度 1，それ以外の質問を適合度 0 とする．
- 多値適合性: もっとも多くのワーカから最も共感できると判断された質問を適合度 2，それ以外で少なくとも 2 人以上から最も共感できると判断された質問を適合度 1，それ以外を適合度 0 とする．

5.4.4 結　　果

前節で述べた 2 種類の適合度判定方法での，20 件のクエリに対する各手法の検索精度の平均を表 6 と表 7 にそれぞれ示す．

表 6 と表 7 をみると，2 値適合性の評価方法では，状況類似判定のファインチューニングを行った BERT が最も高い nDCG@1 と nDCG@3 を，多値適合性の評価方法では TF-IDF と状況類似判定のファインチューニングを行った BERT が最も高い nDCG@1 となっていることが分かる．この結果は，状況の類似性に着目した類似文を用いてファインチューニングを行う提案手法が，共感できる悩みの検索に有効であることを示唆しているといえる．

より詳しく各手法の検索精度を分析するため，それぞれの手法でのランキング結果を分析した．まず，最も多くのワーカが提案手法である，状況類似判定のファインチューニングを行った BERT から検索された質問を選択したケースを分析した．表 8 に状況類似判定のファインチューニングを行った BERT が 1 位に出力した結果を示す．この質問と悩み文は「気力が出なくて，料理や子供の世話が出来ない」という点で類似していると考えられ，実際，ワーカがこの質問を選択した根拠も「夕飯を作る気力がない」や「育児に追われて家事がままならない」といったことが記述されていた．この質問は TF-IDF や Okapi BM25，事前学習のみの BERT では上位 3 件以内にランキングされておらず，BERT を状況的な類似性に注目してファインチューニングすることで，状況の類似性を捉えることができた 1 例であると考えられる．

次に，状況類似判定のファインチューニングを行った BERT に基づく検索が有効に働かなかった例を分析する．この手法で検索した質問を最も共感できると選んだワーカが 1 人以下のクエリは 20 件中 7 件あった．それらのクエリがどのようなクエリであったかを分析したところ，7 件中 6 件は訓練データと大きく異なる悩みを含む文であった．大規模な日本語コーパスで学習された，事前学習済みの BERT をファインチューニングすることで，訓練データとあまり類似しないような悩みに対しても類似質問検索が有効に働くことを期待したが，今回の実験ではそのような結果は得られなかった．今回はファインチューニングに用いた悩み文が 40 件と少ないため，今後学習に用いる悩み文をより増加させることで精度がどの程度向上するのかを検証することが必要であると考えられる．

6 ま と め

本研究では，悩み相談者が投稿した悩みを含む文をクエリとして，CQA コンテンツからこの悩み相談者が共感できると考えられる質問を検索する問題に取り組んだ．本研究は，悩みの状況が類似することが重要であるという仮設を立て，悩みの状況に着目した類似文検索手法を提案した．本問題に取り組むため，CQA コンテンツから抽出した悩みと状況が類似する悩みをクラウドソーシングで収集した．このデータを用いて，BERT のファインチューニングを行うことで，悩みの状況に着目したモデルを構築した．類似文検索手法の比較手法として TF-IDF，Okapi BM25，事前学習のみの BERT を用いて，提案手法を評価した．20 件の悩み文を用いて CQA コンテンツから類似質問を検索する精度を評価した結果，状況の類似性に着目した類似文判定のファインチューニングを行った BERT が最も高い nDCG@1，nDCG@3 を達成した．

謝　　辞

本研究の一部は JSPS 科学研究費助成事業 JP16H02906，JP18H03494，JP17H00762，JP18H03243 による助成を受けたものです．また，本研究では，国立情報学研究所の IDR データセット提供サービスによりヤフー株式会社から提供を受けた「Yahoo! 知恵袋データ（第 3 版）」を利用しました．ここに記して謝意を表します．

表 8　BERT（状況類似判定）が有効に働いたと考えられる例．図の出力は提案手法が 1 位にランキングした質問中の，クエリと最も類似度が高いと判断された文．

クエリ	出力（1 位）
毎日子供に振り回され、夕方にはヘトヘト疲れ果てて気力もないため、ろくに夕飯作れてません。	たまに夕飯も作る気力もなく子供達をお風呂に入れてやれない日もあります。

文　　献

[1] Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. Cross-domain modeling of sentence-level evidence for document retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3490–3496, 2019.

[2] Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. Learning to retrieve reasoning paths over wikipedia graph for question answering. *ArXiv*, Vol. abs/1911.10470, , 2019.

[3] H. Campbell, Marie Phaneuf, and Karen Deane. Cancer peer support programs - Do they work? *Patient education and counseling*, Vol. 55, pp. 3–15, 2004.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 17th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[5] Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. CEDR: Contextualized embeddings for document ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1101–1104, 2019.

[6] Wataru Sakata, Tomohide Shibata, Ribeka Tanaka, and Sadao Kurohashi. FAQ retrieval using query-question similarity and BERT-based query-answer relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1113–1116, 2019.

[7] Carolyn E. Schwartz and Rabbi Meir Sendor. Helping others helps oneself: response shift effects in peer support. *Social Science & Medicine*, Vol. 48, No. 11, pp. 1563–1575, 1999.

[8] Miriam Stewart, Karina Davidson, Darlene Meade, Alexandra Hirth, and Patty Weld-Viscount. Group support for couple coping with a cardiac condition. *Journal of advanced nursing*, Vol. 33, pp. 190–199, 2001.

[9] 岡野禎治, 岡崎祐士. メンタルヘルスケア. 産科と婦人科, Vol. 14, pp. 37–56, 2007.

[10] 玉木敦子. 産後のメンタルヘルスとサポートの実態. 兵庫県立大学看護学部・地域ケア開発研究所紀要, Vol. 14, pp. 37–56, 2000.

[11] 高村寿子, 松本清一. 性の自己決定能力を育てるピアカウンセリング. 小学館, 1999.

[12] 小野美穂, 高山智子, 草野恵美子, 川田智恵子. 病者のピア・サポートの実態と精神的健康との関連. 日本看護科学会誌, Vol. 27, No. 4, pp. 23–32, 2007.